1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

- Below are the variables having any kind of trend with the target variable
    - Season
        - Summer and fall have more rentals
    - Month
        - Follows a similar trend to season
    - Yr
        - 2019 has more rental as compared to 2018
    - Weathersit
        - Clear weather results in more rental
    - Throughout the week the rentals are similar
    - Working days and holidays do not have much impact

2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans.

By dropping the first columns from the dummy variable we remove one complexity from the model, without having any impact on the quality of the model

- All the categories are represented by n-1 dummy variable, n being number of categories, and the missing dummy variable category being represented by all zeroes in the rest.
- Due to the above-mentioned point, the model is can be considered to be defaulted for the missing categorical dummy variable

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans.

- temp.
    - atemp also has similar corelation

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

- Using residual analysis.
    - The mean of error terms is zero (distribution plot for error plot)
    - Error terms are normally distributed (distribution plot for error plot)
    - Error terms are independent of each other (scatter plot between y-train and error)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans.

- Temperature, one degree increases in temperature increase the demand by 0.441 if all other variable remain the same.
- Rain, when all other variable remains the same, the demand decreases by .246 if it is raining as compared to not raining scenario

- Year, as compared to 2018 the rental are 0.232 more for 2019

1. Explain the linear regression algorithm in detail.

Ans.

2. Explain the Anscombe's quartet in detail.

Ans
.

3. What is Pearson's R?

Ans.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.

- Scaling is the technique used to bring the relative values between two variables down. For e.g., a variable has a range 1-10 and another has a range pf 1000-100000.
- Gradient descent algorithm may take a long time to identify the minima of the cost function if the scale of variable is too different. It also makes the interpretation of the resulting model difficult
- In case of normalizes scaling the variable is squeezed between 0 and 1, in standardized scaling the scaling happens such that the mean becomes zero and standard deviation 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans.